DOCUMENT RESUME

ED 296 010 TM 011 940

AUTHOR Forster, Fred; Karr, Chad
TITLE Comparability of Test Scores for the Same Individual: Implications for Vertical Equating.
PUB DATE Apr 88
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Elementary Education; *Elementary School Students; *Equated Scores; *Latent Trait Theory; Scaling; Scores; Standardized Tests; Statistical Analysis; Testing Programs
IDENTIFIERS Rasch Scaled Scores; Survey of Basic Skills; Tests of Individual Performance; *Vertical Equating

ABSTRACT

A method for equating test scores between two standardized achievement testing programs was developed. The first test was the Survey of Basic Skills (SBS) published by Science Research Associates. The second was the Tests of Individual Performance (TIP) of the Portland Public Schools in Oregon. Scores reported in Rasch units (RIT) from the TIP administrations to a total of 1,918 students in the Tigard Oregon schools, grades 3 to 6, were matched to SBS item responses; and SBS tests were scaled (SCS scores) and linked to RIT scales using the fixed parameter method. There was a high level of consistency for the SCS/RIT conversion tables. The TIP RIT and SBS SCS scores provided comparable measures of reading and language achievement for these students. This study supports the possibility of vertically equating standardized test series in sequential subjects. Five figures are provided. (SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED296010

# COMPARABILITY OF TEST SCORES FOR THE SAME INDIVIDUAL: IMPLICATIONS FOR VERTICAL EQUATING

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

CHAD KARR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Fred Forster, PhD

Microsystems for Education and Business, Inc.
395 SW 6th Street
Gresham, OR 97080

Chad Karr, PhD

Department of Psychology
Portland State University
Portland, OR 97207

Presented at the annual meeting of the
American Educational Research Association
April 1988
New Orleans, Louisiana

TM011940

BEST COPY AVAILABLE

## Comparability of Test Scores for the Same Individual: Implications for Vertical Equating

### Overview

The main function of a standardized achievement test in a school district is providing a standard for comparing the performance of students with one another or with the performance of students in other districts. In addition, the information provided by these tests is often used to place students in instructional groups or special programs such as Talented and Gifted or Chapter 1. Typically, the district averages on the standardized test are the primary evidence used by school administrators to monitor the overall performance of the students in their schools.

Two concepts, central to use of standardized achievement tests data, are often confused. These concepts are norms and scales. Norms make it possible to compare a test score obtained by a student to the performance of other groups of students. Usually these norms are developed by giving the test to hundreds or thousands of students and then using the distributions to develop transformed scores such as the percentile, the standard score, the normal curve equivalent, or the grade equivalent scores. The large sample on which the test is normed is usually designated as the "national norm group" and arguments are presented to show that it is representative of the national population.

The concept of a scale (theoretically) is quite different from that of a norm. Allen and Yen (1979, p.179) define a scale as "an organized set of measurements, all of which measure one property or trait." In the present context it is sensible to limit the term scale to designate the "equal interval" property of test scores which make it possible to calculate meaningful averages and distributions. In the past, nearly all efforts to develop a scale were derived from norm distributions, thus confounding the difference between these concepts. But the advent of so-called latent trait models make it possible to approach the issue of scaling directly.

A norm has the limitation that it is anchored to a particular group given the test at a particular time. Therefore, the appropriateness of the scale for a given situation depends on the careful appraisal of the match between the group on which the norms were developed and the group currently being tested. For example, the norms provided by most test publishers are of little use in identifying whether a specific group of Chapter 1 (remedial) students are moving ahead at an acceptable rate. In contrast, an equal interval scale gives meaning to a score <u>independent of the group on which the scale was developed</u>. For example, we accept the measurements provided by a ruler or tape measure without thought to the actual objects used by the manufacturer to create the length scale they represent. Similarly, we feel confident in using these scales to measure different objects so long as they validly represent the "trait"

in which we are interested, e.g., the same ruler can measure a child's height, the size of a package, etc.

Perhaps the most important distinction between an equal interval scale and a norm comparison is the "additivity" of scores derived from the scale. Though it is common practice in some school districts to average percentiles or grade equivalent scores, there is no justification in the nature of these scores for such a practice and no meaningful way to interpret the resulting averages. On the other hand, scores derived from an equal interval scale represent, in some standard way, equal units which can be averaged, squared, etc. to develop the statistics we usually use to describe a distribution.

The measurement models derived from item response theory (IRT) have made it possible to approach the development of true equal interval scales and to evaluate the degree to which that development has been successful. Major test publishers and some large school districts now use this new method of test development. However, this new method and its one, two, and three parameter variations have important limitations. The exploration of those limitations is the focus this research paper.

## Review of the Literature

A frequent focus of research articles in recent years, especially those appearing in the Journal of Educational Measurement, has been the actual success of IRT models in developing true equal interval scales. For example, researchers have identified specific situations in which the scales developed using these models do not provide satisfactory measurement quality. In their thorough review of studies related to "horizontal and vertical equating," i.e., the attempt to link scales using tests of different difficulty levels, Skaggs and Lissitz (1986) reviewed 28 studies. Based on this review, they concluded that, for horizontal equating, "IRT equating seems to produce better results than conventional methods" and later, for vertical equating, "The Rasch and linear models produce inadequate results." and "Evidence suggests that three parameter model equating provides better results than the Rasch model and linear equating, but it has not been consistently superior to equipercentile equating."

The equating of different forms of published standardized tests and tests from different publishers, has been addressed by several articles. The most notable of these efforts was the Anchor Test Study (ATS) (Loret, Seder, Bianchini, & Vale, 1974) planned to provide a linkage between the norm scores of the major published standardized tests. The data from this study was used by Slinde and Linn (1977) to evaluate the comparability of norms resulting from equating these tests using the equipercentile method. They concluded that "Based on a logical analysis, as well as empirical comparisons using Anchor Test Study results, it must be concluded that the vertical equipercentile equating of

existing tests, when using the base-test approach, leaves something to be desired."

The data from the ATS were re-analyzed by Rentz and Bashaw (1975, 1977) and they concluded that the ability estimates determined for their new scale based on this equating "...are quite stable and invariant with respect to replications of the estimation process. They would seem to support the contention that there is sufficient "Scoring-Table" invariance to warrant confidence in the equating results...." Slinde and Linn (1979) also reanalyzed the ATS data using the Rasch model and a cross-validation design. They concluded that "The results of this study indicate that the Rasch model has greater utility for the purpose of vertical equating than was apparent in the previous Slinde and Linn studies.... The results of this study indicate that there is reason to be optimistic about using the Rasch model to scale existing test data in order that objective measurement can be achieved."

Another study reported by Slinde and Linn (1978), however, did not yield results nearly as positive. Dividing items on the College Entrance Board's Mathematics Achievement Test into easy and difficult subtests, they concluded that "For the math data analyzed in this study, the Rasch model did not provide a satisfactory means of vertical equating.".

In a more negative vein, Divgi (1986), in an article titled "Does the Rasch Model Really Work for Multiple Choice Items? Not if You Look Closely," used the ATS data to analyze the fit of the standardized test items to the Rasch model. He concluded that "Of the total of 1,942 items in the 21 test-grade combinations, 1,326 (68.3%) and 329 (16.9%) are misfitted according to MSq and MSt, respectively."

A study reported by Philips (1986) attempted to determine the effect of removing "misfitting" data during the equating process. The results of this study using fourth and eighth grade on the 3-R's achievement test battery led Philips to conclude "...(a) that the Rasch and recalibrated Rasch (misfitting persons deleted) methods of equating provide similar definitions of score-to-score equivalence when the percentage of misfitting persons deleted is small; and (b) that the two Rasch methods provide somewhat different equated grade development scores than the equipercentile method."

In a study comparing the results for the reading comprehension subtest of the Iowa Tests of Basic Skills for students from ages 9 to 14 (grades 4 to 9), Guskey (1981) concluded that "Throughout the range of test scores, the Rasch ability estimates were found to be at least as precise as the grade-equivalent estimates and were perhaps the more precise estimate where discrepancies were found to exist."

And in the final study to be reviewed here, Beard and Pettie (1979) used the third and fifth grade data from the communications and mathematics skills tests of the Florida Educational Assessment to compare linear and Rasch equating. They concluded "From the results of this report, differences

between the linear and Rasch equating were so small as to appear inconsequential."

Comments on the Literature

As this review indicates, there is little (if any) consensus in the literature on vertical equating, especially as it applies to the Rasch model. We are appreciative of the efforts of Skaggs and Lissitz to even tackle the problem of trying to review these contradictory studies and put them in perspective. But we believe that they, like many of the studies they report, have missed some of the fundamental reasons that have led to the unevenness of the results and conclusions of these studies.

First, the question should be addressed, "When is it reasonable to equate tests." The fundamental problem in many of these studies is that this basic question is ignored when the measurement model and equating methodology are chosen. For example, it seems reasonable to take the position that it is only reasonable to equate tests that, in some logical sense are "equal." This criterion, then, excludes studies which would attempt to equate tests in different subjects, tests covering different subject matter, tests covering material on which students have not been instructed, etc. Similarly, it is the responsibility of the investigator to show the degree to which equating in "non-sequential" subjects such as science or social studies is dependent on strict specifications regarding previous instruction, etc.

Second, the question should be addressed, "For which students is it reasonable to equate tests." For example, it seems reasonable to expect that the results from two tests can be equated only if there is some population for whom both tests are appropriate. If one of the tests is well matched to a group and a second is either much higher or much lower, it should be expected that the results of equating will be confounded with the response of students to a test that is either much too hard or much too easy. This consideration also applies to students who score near or below the "chance" level on a multiple choice test. Attempting to equate these scores or to use student scores in this range open the study to threats to its validity. Normally, in standardized testing in the public schools, students are discouraged from random marking, and a score in this range is an indication that either the test was much too difficult for the student or that the administration was faulty (the student was off in marking the answers, was ill, etc.). Therefore, much of the variation that might be attributed to "guessing," actually is due to invalid administration or test taking behavior where the most reasonable approach would be to drop the data from further consideration. In the case of the studies using data drawn from examinations for college entrance, the dynamics of guessing can be expected to shift dramatically. Because of the impact of these results on the student and the resulting pressure, it is much more reasonable to assign a role to random marking, at least

for students at the low end of the distribution. In either of these cases, if the investigator chooses to use these chance data in the equating, the burden of proof rests with him or her to show that it is a valid response to the test rather than an idiosyncratic response to an unreasonable testing situation or a problem in administration. In the absence of this proof, it seems reasonable to exclude these data not only from the equating analysis, but also from any comparison of scores produced by the equating.

There are other factors which influence the validity of test results and need to concern the investigator. These include the mind-set and attitude of both the teacher and the student toward the test being taken, violations in administration such as allowing too much or too little time, coaching students or teaching specific items on the test, and errors in scoring where light marks or incomplete erasures are not checked carefully.

After these issues have been handled, it seems reasonable to select an appropriate measurement model to do the equating analysis. For example, a latent trait model seems an unlikely choice for the analysis of data for an analogy test in which some items are presented in a sequence to test learning that is occurring, i.e., the test author is purposely violating the assumption of item independence. The application of these models to subjects which are not "sequence-bound" also raises important questions about the determination of results. For example, most concepts in a science or social studies course can be taught using a variety of content. It is reasonable to suppose that, in these subjects, at least two dimensions must be handled (content and concept) by the measurement model, and either one or all of the dimensions may not be strictly sequential. In fact, the investigator must provide a rationale for supposing that the trait being measured is sequential and how other dimensions (such as content, etc.) will be handled to ensure the applicability of the model. If the scientific method can be taught using either mechanics (physics) or acid-base reduction (chemistry), what corollary procedures (or instruction) are required before the current test could be considered a valid measure of that trait. The point of this discussion is that many of the studies reported by Skaggs and Lissitz (1986) should not be expected to provide consistent equating results since two or more dimensions are likely to be involved and evidence is not available on how these dimensions have been conceptualized and controlled.

We believe the reported studies on vertical equating suffer because the authors do not agree on the criterion that would show that the tests had, in some sense, been equated. If we try to emulate the criteria used for instrumentation and measurement in the so-called "hard sciences," we must move toward criteria which are (insofar as possible) independent of the particular sample in the equating study. Using this argument, the mean square fit (or fit mean square) statistic seems wholly inadequate. Proceeding beyond the arguments of parameter distributions, etc., it can be shown (both theoretically and empirically) that this statistic

suffers from the same problems of interpretation as the t, F and Chi-square tests for large samples. As the sample size increases, the fit statistics draw us to the tautology that the model does not match our data perfectly. But this is not the same as addressing the critical question of "How well does the measurement model match the data." As with the physical sciences, the criteria used must decide, not whether our model is perfectly accurate, but whether it is "close enough" or more accurate than an alternative model. Therefore, it seems reasonable to use the reproducibility of the estimates for parameters or scores as the primary criteria for an equating study. And it may even more essential to use a planned cross-validation comparison from different data sources to provide a rationale for statistical independence.

Finally, it should be reaffirmed that a measurement model cannot make up for logical or methodological errors. For example, if half the students in a sample have been encouraged to random mark questions they are unable to answer and half have not, a source of error will have been introduced that will affect the outcome of equating. Similarly, including students with scores below chance (on multiple choice examinations), using data drawn when students are under external pressure (college entrance examinations), or using data from students who have not been instructed in the content of the test introduce serious sources of error which must handled before equating can be meaningful. Most important, there must be a rationale for the use of the model. It is unfortunate that jargon such as "person-free" and "item-free" have obscured "common sense" in the analysis of test data. For example, Divgi's (1986) application of a mean square fit criterion in the face of evidence that it required the discarding of at least 30% of a pool of items with impeccable statistical credentials, does not seem to be professionally responsible. It would have been much more reasonable to assume that the fit statistics used in the study had failed and to look for a more relevant criterion on which to base the data analysis. Similarly, equating studies which do not include a cross-validation component, do not inspire much confidence. With the many potential threats to validity arising from problems in preparing the students, administration of the tests, data handling, scanning, etc., any non-replicated result must be considered in doubt until a replication is available. Most important, there should be a recognition that the success of an equating study rests on the quality of the data being analyzed. And the quality of the test data depends on the instructional preparation provided to students, the cooperation of teachers and students, the conditions and proctoring during the administration of the test, the care spent checking for light marks or multiple marks as part of scanning, etc. Where these conditions are not met (or not known), the usefulness of the equating study is limited severely.

## Purpose for the Study

The purpose for this study was to provide a method for equating test scores between two standardized achievement testing programs. The first program was the Survey of Basic Skills (SBS) test battery, published by Science Research Associates. The second program was the Tests of Individual Performance (TIP) published by Portland Public Schools.

## Methodology

Students in the Tigard (Oregon) Public Schools, grades 3 to 6, took the TIP tests in fall 1985 and spring 1986. (There were 480 students in grade 3, 481 students in grade 4, 500 students in grade 5 and 457 students in grade 6.)

The TIP testing program uses a "level" approach where each student is given a test matched to his/her achievement level at the time of testing. The scores from the TIP program are reported in Rasch units (RIT's) based on scales in reading, mathematics, and language which have been used by school districts in Washington and Oregon. The TIP programs were coordinated so test results were returned to the classroom the week following testing. Students who scored outside the "valid range" of the test (below 25% or above 90%) was retested with a more appropriate level (either higher or lower).

The SBS testing program uses a grade level approach where all students in a given grade take the same examination. The scores provided by the SBS program include the SCS, a Rasch scaled score, and the GSV, a growth scaled value score. The test results from the SBS program were returned to the schools approximately six weeks after testing. Students scoring outside the valid range of the test were not retested. Following the publisher's recommendation, students in grade 3 took SBS level 23, students in grades 4 and 5 took SBS level 34 and students in grade 6 took SBS level 35.

The TIP testing was administered in October 1985 and October 1986. (The TIP program had been in place since 1983.) The SBS testing was administered in May 1986. (This was the first administration of the SBS program.)

The RIT scores from both TIP administrations were matched to the SBS item responses using the district student identification numbers. The SBS tests were scaled and linked to the RIT scales using the Fixed Parameter method (Forster & Karr, 1986) which provided a RIT score for each possible raw score on the SBS tests. Test scores below 25% and above 90% were dropped from the scaling analysis. The resulting data file contained one or two TIP RIT scores, a SBS RIT score, a SBS SCS score and a SBS GSV score for each student.

## Results

The first question to be addressed was the accuracy of scaling of the SBS tests to the RIT scale. Since the SBS tests had been previously linked by the publisher (as described in their manual), SCS scores were available to provide a link between raw scores for levels 23, 34 and 35. An independent link was available based on the Fixed Parameter analysis for the students in this study which linked the raw scores on the three SBS tests to the TIP RIT scales and adjusted for the time differential between TIP and SBS testing (Forster & Karr, 1986). The plot of SCS scores against RIT scores provided a conversion between the scales. The degree to which the three plots overlap (or diverge) indicates the stability of this conversion.

As shown in Figures 1 and 2, the three plots of SCS against RIT formed virtually a straight line, indicating an extremely high level of consistency for the SCS/RIT conversion tables, for both reading and language.

The strength of this relationship was surprising, because of the skewed distribution of the SBS scores. The distribution shown in Figure 3, for level 34 (4th and 5th grade) reading raw scores, is typical of the distributions for both subjects and all three tests. In a previous study (Forster, 1987) it had been found that negative skewness (much more severe than in the current data) had adversely affected the scaling. The consistency of the SCS/TIP plots is evidence that the less severe negative skewness of these scores did not adversely affect the Fixed Parameter scaling of the SBS tests.

The second question to be addressed by this study was the comparability of the scores from the two testing programs. Figures 4 and 5 show scatterplots for fall 1985 TIP reading to spring 1986 SBS reading and fall 1986 TIP reading scores for the 4th and 5th graders. The correlations were comparable for the two scatterplots. Therefore, it appeared that the TIP RIT and SBS SCS scores provided comparable measures of reading and language achievement for these students.

## Conclusions

The literature on vertical equating is riddled with inconsistent findings. Part of this inconsistency is due to a focus on the modeling aspects of the equating rather than the resulting scales. Part is probably due to the lack of an independent criterion for the accuracy of the scaling. And part is likely to due to the choice of subjects in which instruction does not follow a roughly linear sequence as it does in reading, general mathematics, and language usage.

The present study supports the possibility of vertically equating standardized test series in sequential subjects from data drawn as part of carefully administered district programs.

## Directions for Future Research

This study is part of a broader research effort by the authors to establish the boundary conditions under which a latent trait model is likely to provide a satisfactory equating between two tests used in elementary or high school assessment.

In the vein of this study, a replication is planned using grades 3 to 11 with the SBS test series. A second replication is underway equating a different published standardized test series to the TIP RIT scales.

Beyond this, a study is in progress to explore the consistency between the TIP general mathematics scale and scales for algebra, geometry and analysis. Finally, studies have been planned to explore the effect of the "ceiling" on tests (such as competency examinations) on the resulting equating to a typical standardized test or another competency examination.

## REFERENCES


Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole.

Beard, J. C., & Pettie, A. L. (1979, April). A comparison of linear and Rasch equating results for basic skills assessment tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. Journal of Educational Measurement, 23, 283-298.

Forster, F. (1987, April). Riding the Rasch tiger: Part 1: Laying the item bank foundation. Paper presented at the National Association of Test Directors Symposium held during the annual meeting of the National Council on Measurement in Education, Washington, DC.

Forster, F. & Karr, C. (1987, April). The accuracy of the fixed parameter method of item bank calibration. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Guskey, T. R. (1981). Comparison of a Rasch model scale and the grade-equivalent scale for vertical equating of test scores. Applied Psychological Measurement, 5, 187-201.

Loret, P. G., Seder, A., Bianchini, J. C., & Vale, C. (1974). Anchor test study final report: Project report, Vols. 1-30. Princeton NJ: Educational Testing Service.

Philips, S. E. (1986). The effects of the deletion of misfitting persons on vertical equating via the Rasch model. Journal of Educational Measurement, 23, 107-118.

Rentz, R. R., & Bashaw, W. L. (1977). The national reference scale for reading: An application of the Rasch model. Journal of Educational Measurement, 14, 161-179.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. Review of Educational Research, 56, 495-529.

Slinde, J. A., & Linn, R. L. (1977). Vertically equated tests: Fact or phantom? Journal of Educational Measurement, 14, 23-32.

Slinde, J. A., & Linn, R. L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 15, 23-35.

Slinde, J. A., & Linn, R. L. (1979). A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. Journal of Educational Measurement, 16, 159-165.

Figure 1

PLOT OF TIP READING RIT VS. SCS SCORES
BASED ON THE FIXED PARAMETER SCALING

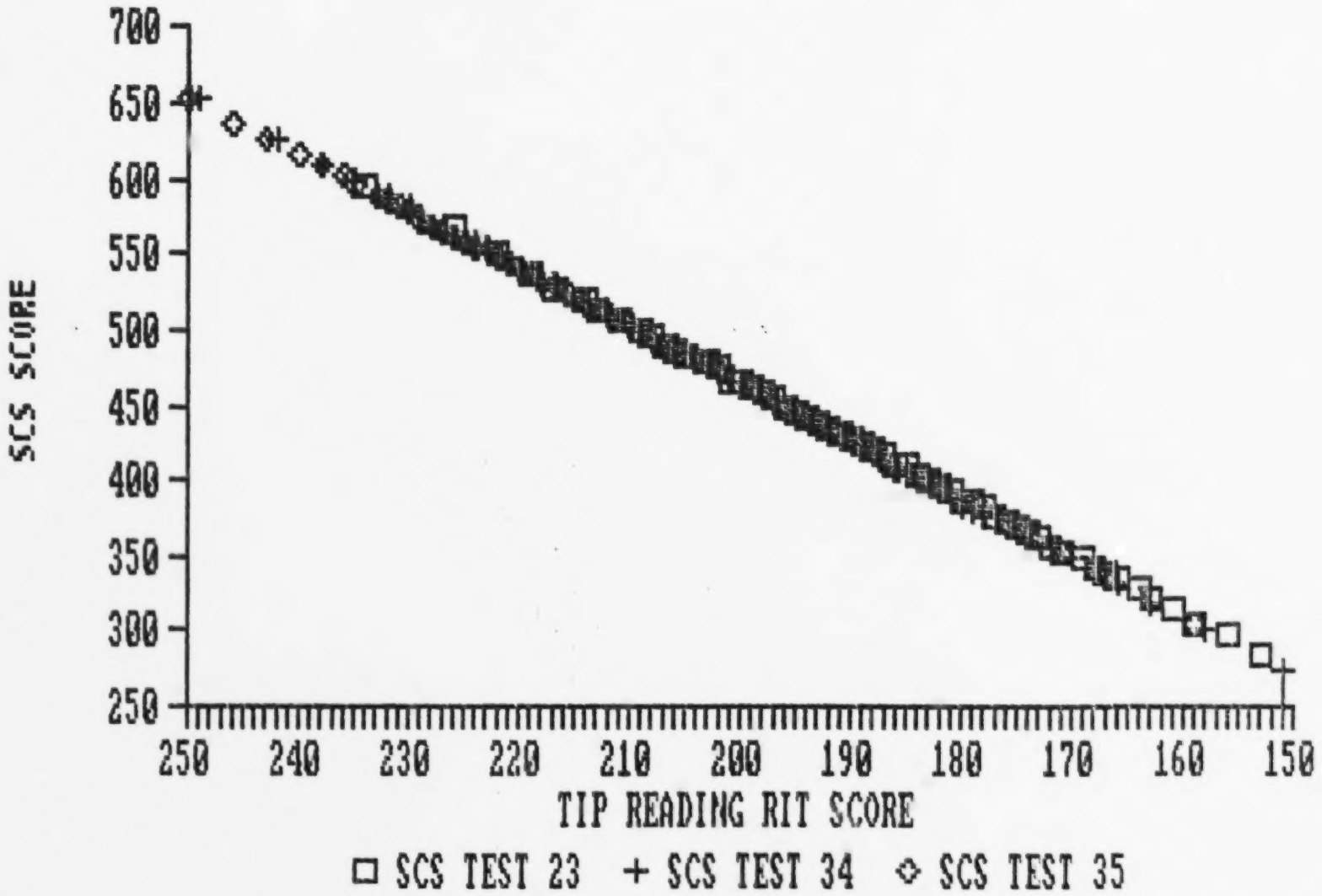
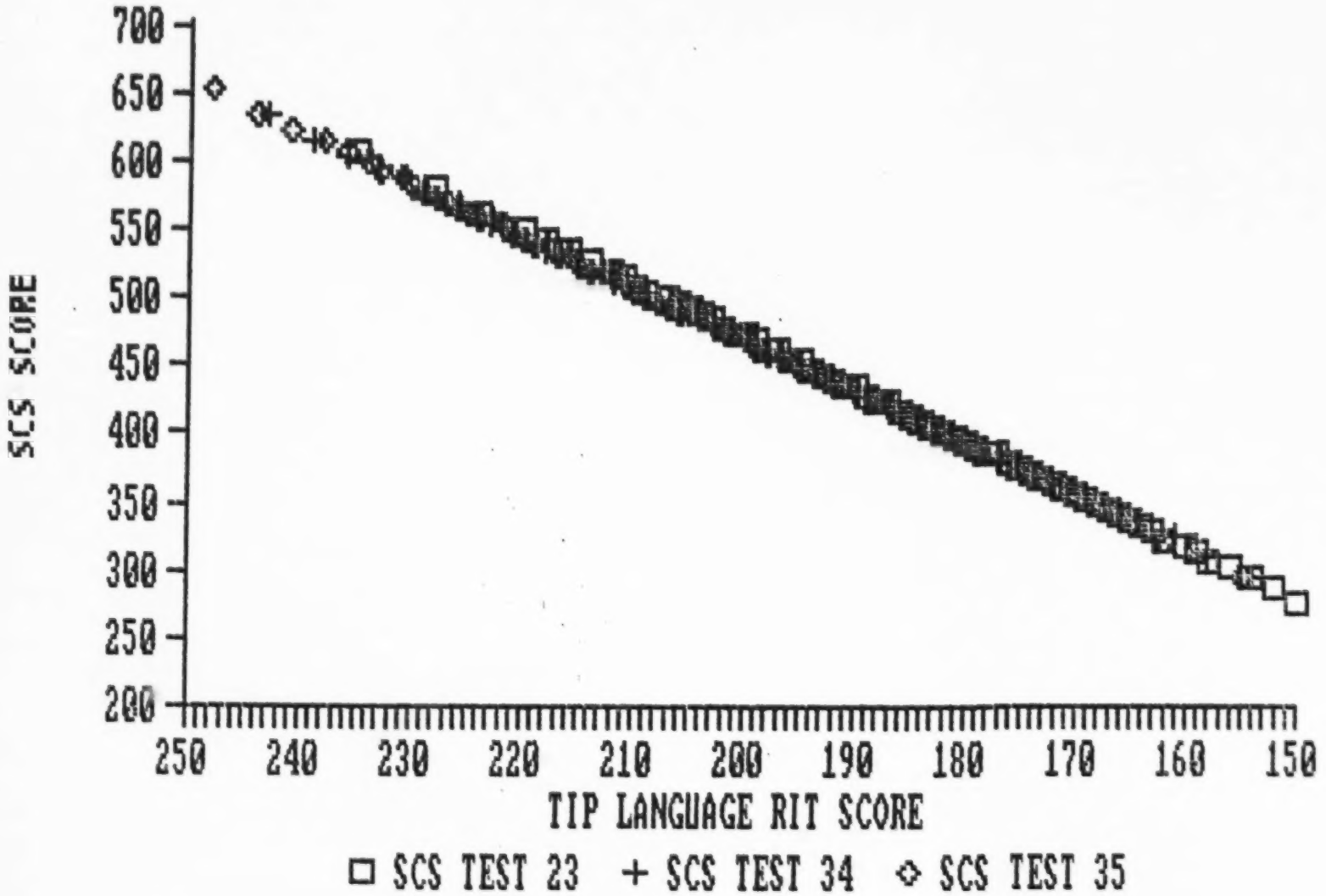


Figure 2

PLOT OF TIP LANGUAGE RIT VS. SCS SCORES
BASED ON THE FIXED PARAMETER SCALING



Distribution of Raw Scores for the SBS Reading Test
Level 34 (Grades 4 and 5)

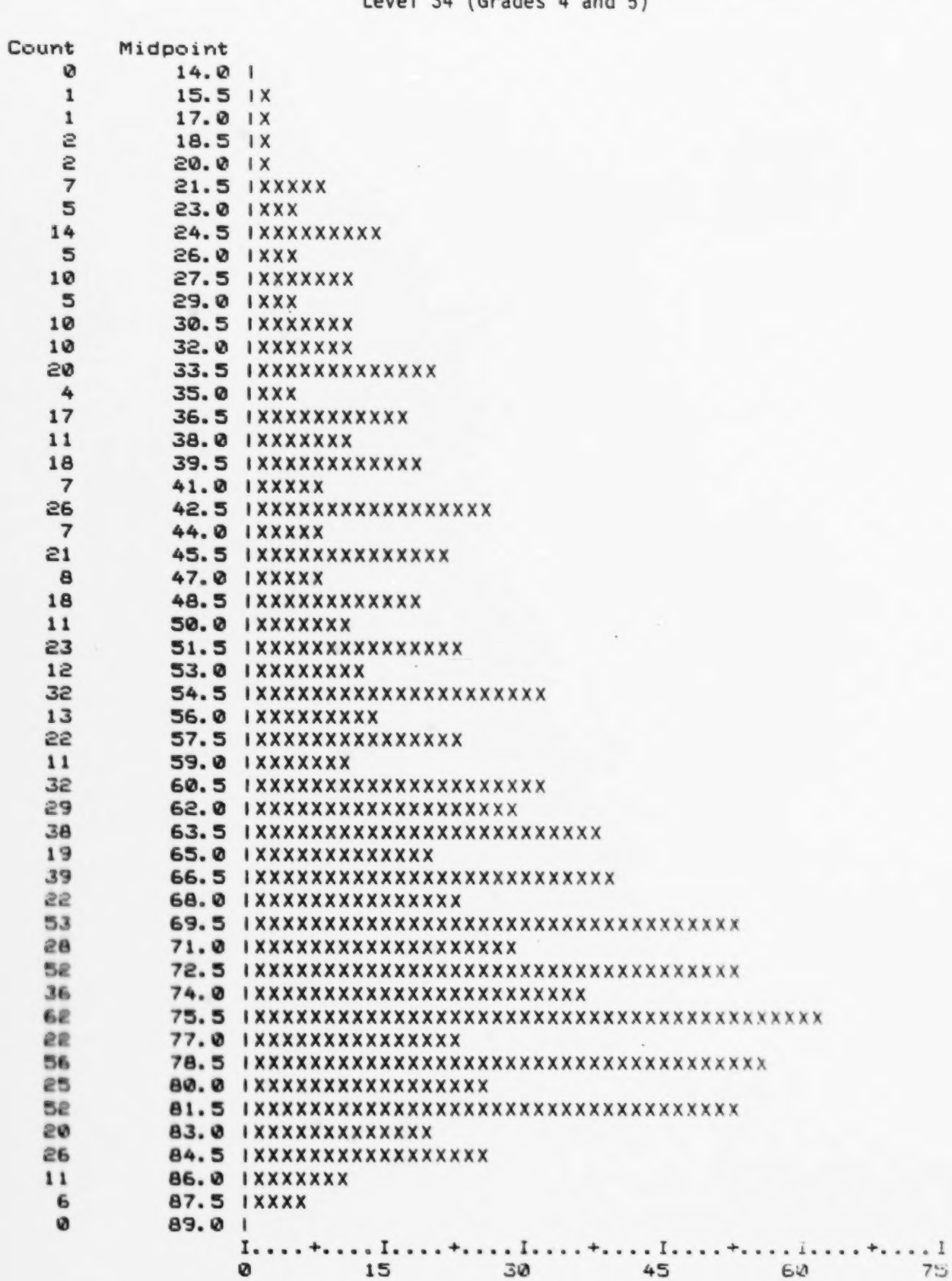
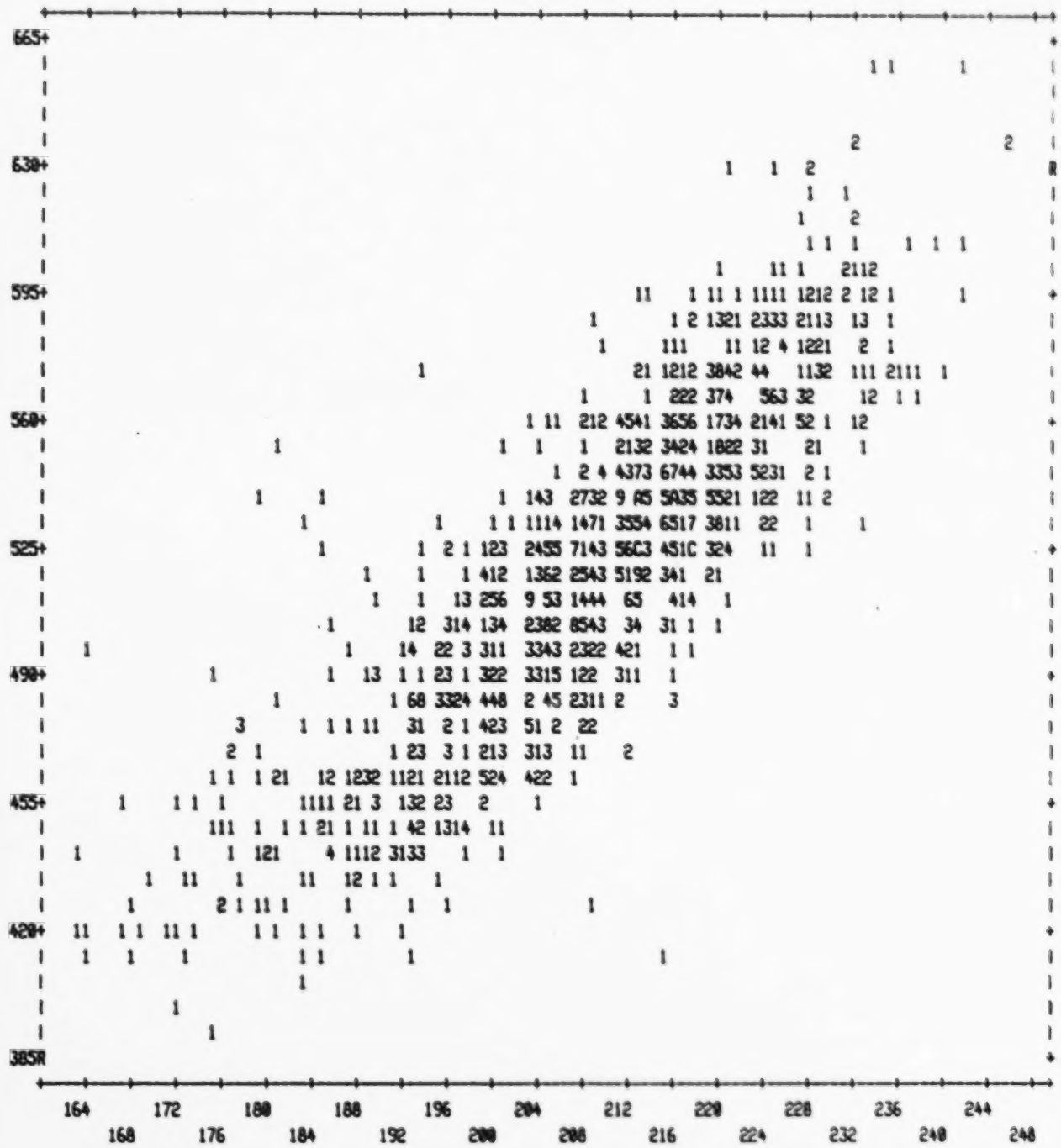


FIGURE 4

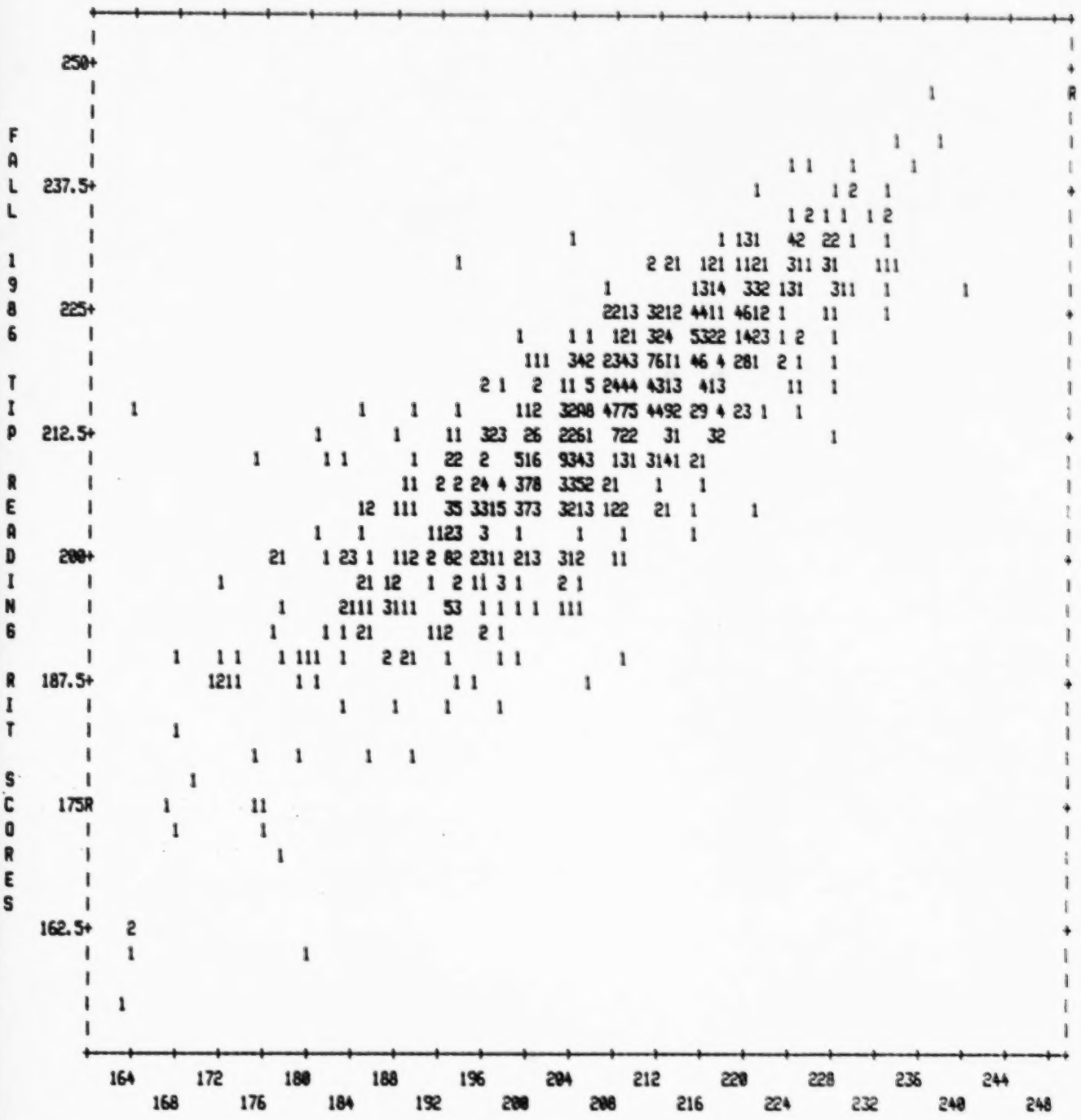
Scatterplot of Fall 1985 TIP Reading RIT Scores Against
Spring 1986 SBS Reading SCS Scores




1287 cases plotted. Regression statistics of SBSR_SCS on TIPR_RIT:
Correlation .84313 R Squared .71087 S.E. of Est 24.24864 2-tailed Sig. .0000
Intercept(S.E.) -53.77235( 10.23229) Slope(S.E.) 2.74613( .04886)

FIGURE 5

Scatterplot of Fall 1985 TIP Reading RIT Scores Against
Fall 1986 TIP Reading RIT Scores




799 cases plotted. Regression statistics of XIPR_RIT on TIPR_RIT:
Correlation .84166 R Squared .70839 S.E. of Est 6.90945 2-tailed Sig. .0000
Intercept(S.E.) 46.96285( 3.76676) Slope(S.E.) .80318( .01825)